CENTER FOR DEMOCRACY & TECHNOLOGY

August 9, 2022

The Center for Democracy & Technology welcomes the opportunity to provide comments on case 2022-007-IG-MR regarding the takedown of an Instagram post following a request from UK law enforcement alleging that the post may contribute to offline violence.

A post from an account promoting UK drill music was taken down by Instagram following a request from local law enforcement. The post featured a short clip from a new drill music video from rapper Chinx (OS). The law enforcement authority informed Meta that the post referenced a past shooting and could provoke further violence. It is not clear from the Case Summary whether the authority identified the post to be a violation of local UK law. The post was then reviewed by an internal team at Meta and was taken down for violating the company's Violence and Incitement policy, which Meta informed the Board can only be enforced by Meta's internal teams. The user who created the post was notified by Meta both times their content was removed but was not informed that the removals were initiated by a request from UK law enforcement.

This practice of law enforcement flagging content for companies to review against their Terms of Service (TOS) is part of a growing arsenal state authorities use to pursue content removal from online services. Law enforcement authorities may do this informally or through formal structures, such as Internal Referral Units (IRU).[1] This practice raises a number of human rights, transparency, and due process related concerns.

Typically, if law enforcement authorities seek to have content removed from an online service, they should obtain a court order from an independent arbiter and comply with procedural requirements designed to protect the speaker's rights.[2] Flagging content against a company's own internal standards, however, is often a far quicker path for law enforcement to get a company to review and remove content than pursuing legal action against the content, and is often one that includes far fewer safeguards for users' rights. Moreover, a notification from law enforcement may come with overt or implicit pressure on the company to comply, even if the company would not otherwise have considered the content a violation of its policies. Thus, the risk for abuse from these kinds of referrals from law enforcement authorities is high.

[1] The first IRU was the UK's Counter-Terrorism Internal Referral Unit set up in 2010 and run by the Metropolitan Police. Since, the European Commission has called on all EU member states to establish national IRUs to combat the spread of extremist content online. (Chang, 2018 p121)

[2] We focus here on content removal as the action sought by law enforcement and taken by Meta in this case, but we also note that the risks of law enforcement interference with speech and access to information, and the need for enhanced transparency from Meta, are also relevant in situations where some other action, such as demonetization or limiting a post's distribution, is taken.

Consequences of law enforcement referrals of TOS violations

Meta's global Community Standards reflect its decisions about what kinds of content it does and does not want to allow on its services. As a result, the Community Standards can and do restrict speech beyond what national laws or international human rights standards would permit state actors to censor. Law enforcement flagging of content against these broad standards, therefore, may result in taking down content that a law enforcement official could not seek to restrict under the law. This ultimately threatens the ability for users to express themselves free from illegitimate government interference.

The UK government's treatment of drill music exemplifies this danger. Drill music is an art form which depicts "morally charged caricatures" of its artists armed with guns, provocative lyrics, and drugs. The head of the UK Crown Prosecution Services, the top crime prosecution service in the country, has called drill music a genre that, "by its nature, is supposed to shock but [one that is] not a crime". Despite the genre being replete with lawful, protected expression, the UK Metropolitan Police have indexed and tracked an extensive list of drill music videos and shared them with platforms like YouTube since 2019, resulting in the takedown of over 100 videos. Law enforcement has thus flagged content against Meta's Community Standards that does not violate UK law.

This is not a dynamic unique to the UK: In 2019, France's IRU, using an application provided by Europol, sent over 550 baseless referrals to Archive.org about URLs on its service, asserting that links such as the collections page for animation and cartoons, a research paper on male infertility, and an archive of a US House of Representatives hearing were all examples of "terrorist propaganda" that should be removed. Insights from Israel's Cyber Unit, the country's IRU, in May 2021 showed that Meta complied with 46% of flags for content removals; because the company would likely face civil or criminal liability for leaving illegal content up under Israeli law, this may mean that upwards of 54% of the notifications from the Cyber Unit related to what Meta believed to be lawful speech.

By flagging content as a TOS violation directly to an online service provider, law enforcement can circumvent established court procedures and bypass substantive human rights law. This can inadvertently turn companies into avenues for state censorship. A flag from a law enforcement authority, whether informal and or through a formalized process like an IRU, can exert significant pressure on social media companies to comply with removal requests even if the content is not unlawful or a TOS violation. Law enforcement authorities carry the weight of expertise—a dedicated counter-terrorism unit will speak with some authority on what amounts to "terrorist propaganda"—and are armed with other levers of state pressure. Particularly in jurisdictions where issues of intermediary liability are handled through notice-and-takedown regimes, the receipt of a notification from law enforcement about a specific piece of content may

jeopardize the provider's safe harbor from liability, because the provider may be considered under the law to now "know" that they are hosting (allegedly) illegal content.

Moreover, with little data available from companies on the nature of requests they receive and how they comply, people outside the companies cannot completely grasp the extent to which law enforcement requests shape the enforcement of terms of service. Procedural transparency, then, is critical to understanding the role law enforcement referrals play in enforcing limits on user speech.

Transparency is also critical to equip individuals with the ability to assert their fundamental rights against the state. A 2021 ruling from Israel's High Court concluded that, in order to have standing to bring a court challenge to state action against their speech, users need to know whether and when state actors flag their content to technology companies for review. In that case, the plaintiffs, two non-governmental organizations, were unable to prove that their posts were removed from Facebook after, and *because of,* a flag from the local Cyber Unit, because they only received notices from Facebook that their posts violated Facebook's content policies. The High Court found that the plaintiffs did not have the evidentiary basis to bring a challenge against the Cyber Unit for its involvement in the removal of their speech from Facebook. This case demonstrates that a lack of transparency from social media companies doesn't just stand in the way of a dispute between a user and the service provider. It can fundamentally deprive individuals of their access to remedy for violations of their rights by the state.

Recommendations

While responsibility for procedural transparency certainly rests with state actors, Meta also has a responsibility to increase transparency into their handling of referrals from law enforcement. Procedural transparency will serve two important purposes: allowing the public, policymakers, and civil society to understand the extent to which law enforcement bodies play a role in subjecting online speech to scrutiny, and providing notice to individuals when their online activity is being scrutinized by the state and thereby providing them a basis to seek a remedy.

Transparency reporting

First, Meta should provide substantially more clarity on the process it employs when it receives a referral from law enforcement. The case summary describes Meta as saying that the UK law enforcement request for content removal was escalated for an internal review by experts at Meta and *not* subject to review by outsourced moderators. Meta explained this was because the Violence and Incitement policies can only be enforced by Meta internal teams. This process raises questions about whether other TOS removal requests from law enforcement are also enforced by internal teams, whether reviewers are aware of who submits the request, and, if so, whether that influences the reviewer's response. As the Oversight Board has recommended in the

past: Meta must formalize a transparency process on how it receives and responds to government requests for content removal and extend this to law enforcement referrals as well.

Second, Meta should disclose when it complies with law enforcement referrals for content removal. The Oversight Board has recommended that "transparency reporting should distinguish government requests that led to removals for violations of the Community Standards from requests that led to removal or geo-blocking for violating local law, in addition to requests that led to no action."

Meta indicates that it plans to implement this recommendation in full and that it has been "in progress" since 14 October 2021. Meta should prioritize implementing this recommendation promptly. An example of another social media company's approach comes from Twitter, which began in 2017 to disclose the number of times it removed posts or accounts as TOS violations following notification by a government request. Meta should also provide information about the formal IRUs (such as the UK CTIRU or the Israeli Cyber Unit) from which it receives referrals, as well as whether any law enforcement agencies function as "trusted flaggers". When governments leverage private terms of service and circumvent legal processes to restrict more speech faster, it's vital that we have a better understanding of the scope and scale of this activity.

User notifications

Lack of notice to individuals when their post is removed following a notification from a state actor deprives them of the opportunity to challenge this state action. Individuals should know when their government is scrutinizing their speech and be able to seek redress against extralegal censorship. As the Israeli High Court case discussed above illustrates, individuals must be able to articulate whether and how their government was targeting their speech for impermissible restriction in order to have a chance to obtain a remedy. Absent significant changes in governments' own approach to these cases, Meta is the only entity that can provide information to its users about this type of government activity.

When users are notified that their post has been removed or their account deactivated, Meta should also disclose:

- When a post was originally brought to Meta's attention by a law enforcement authority or other government actor, and which part of the government the actor is affiliated with;
- Whether the notification was a legal order or a referral for a TOS violation;
- Whether the notification included an assertion that the content violated national law (and if so, which one); and
- How a user may appeal the decision made by Meta.

For more information, contact Emma Llansó, Director, Free Expression Project, ellanso@cdt.org and Aliya Bhatia, Policy Analyst, Free Expression Project, abhatia@cdt.org.